Modern computer networks generate extensive data that can benefit network research, management, and security. This data represents complex interactions among network nodes, services, and users and is fast-evolving, increasingly encrypted, and highly siloed. These characteristics make it difficult to analyze using traditional methods based on predefined rules and signatures. Machine learning (ML) methods show promise in identifying complex patterns and insights in network data [1]. Yet, these methods often face reliability issues in real-world network operations, due to open problems like lack of training data and the enormous variability in input data [2].

Informed by practitioners and measurements, my research is driven by the goal of **enhancing the applicability and reliability of machine learning for networks** through **data-driven methods, robust system design**, and **security analysis**. I take a holistic look at integrating ML into network operations, adapting each phase of the ML lifecycle to fit network requirements. I seek to maximize compatibility and deployability in current environments.

My work focuses on three practical challenges unique to applying data-driven approaches in networking: (1) the need to acquire *diverse* traffic patterns siloed in different network entities, (2) the need for *scalable* platforms that support real-time decision-making for high-throughput data flows, (3) and the need to adapt to *constantly changing* network characteristics and user behaviors. As shown in Fig. 1, at the foundation of my work is the development of **accessible, reliable, and performant machine learning systems for network data analysis**. These systems and frameworks can help break data silos [3, 4, 5] and can merge with other modalities [6]. They are designed to be compatible with existing infrastructures, which handle the scale, heterogeneity, and complexity of modern networks, enabling real-time [7, 8, 9] and adaptive [10, 11] insights for network management and network forecasting. In addition to network management, I employ **network data analysis for critical issues in security and privacy**, addressing challenges such as threat detection [12, 13, 14], and safeguarding online privacy [15] in a highly connected world.

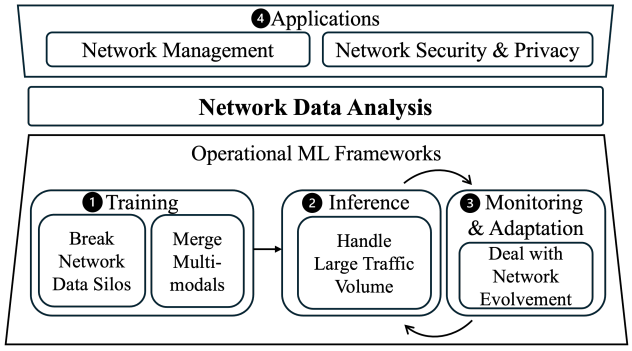

Figure 1: Research Overview

Research impact. The significance and impact of my work has been recognized through acceptance in top-tier conferences (e.g., USENIX Security, SIGMETRICS, CoNEXT, and UbiComp), active industry adoption and media coverage. `LEAF` [10] is being deployed by *Verizon* to provide better forecasts for resource allocation in cellular networks, particularly to combat Concept Drift. The operational workflows of synthetic network traces and generative models such as `NetDiffusion` [4] are integrated into *Rockfish Data* and *Conviva Inc.* My research has also garnered attention from multiple media outlets, including Forbes, The Wall Street Journal, and ACM TechNews.

Dissertation Work

My dissertation examines the holistic integration of machine learning into network operations. My research addresses operational challenges by developing ML frameworks that can seamlessly integrate, scale, and adapt models for real-world networking applications.

❶ **Training: Break Network Silos Using Synthetic Data.** High-quality labeled network trace datasets play a pivotal role in supporting numerous ML applications within the networking domain. However, these datasets are often siloed at different entities and thereby inaccessible due to concerns related to privacy, retention policies, and ongoing maintenance requirements. These limitations significantly hinder full reproducibility in academic work and are also an impediment to collaborations to improve model performance across silos in the industry. Synthetic network traces offer a potential solution to supplement existing datasets, yet current methodologies typically focus on generating aggregated flow statistics or selectively chosen packet attributes. I introduce `NetDiffusion` [3, 4], a framework to generate high-fidelity packet captures (PCAPs) that are protocol-adhered and are compatible with existing hard-

ware. A domain-specific, controlled variant of Stable Diffusion is leveraged. The PCAPs generated by `NetDiffusion` exhibit a higher degree of statistical similarity to real traces and significantly enhance ML model performance in comparison to leading methods. `NetDiffusion` is also able to generate rare instances of network traffic, which augments models on under-performed classes. Furthermore, the generated traces maintain compatibility with widely used traffic analysis tools (e.g., Wireshark), facilitating applications that extend beyond ML tasks. To further enable learning protocol compliance without human intervention, the use of state space models is also proposed to generate packet-level synthetic network traces by framing it as an unsupervised sequence generation problem [5].

Training: Merge Multi-modal Information. Network data serves as a bridge between physical activities and the digital realm. For instance, when users interact with smart devices in a connected environment, these devices generate network traffic that can signal real-world events, aiding in human activity recognition. Other modalities, like video or audio, often correspond to the same activities. However, previous approaches typically require large amounts of "paired" data to effectively leverage cross-modal information, resulting in excessive data re-collection efforts and limiting the reuse of pre-existing models from other modalities.To address this challenge, I propose `AMIR` (Active Multimodal Interaction Recognition) [6], a framework designed to combine video and network data using a meta-learning approach. `AMIR` trains individual models for each modality—video and network activity—and merges their predictions based on uncertainty, enabling a more robust and flexible system. This merging strategy significantly reduces the dependency on "paired" data, where video and network traces are collected simultaneously. `AMIR` achieves up to a 70.83% reduction in paired data requirements while maintaining an 85% F1 score. Additionally, it improves recognition accuracy by 17.76% with the same sample size compared to traditional methods, demonstrating its effectiveness across both controlled lab and real-world home environments.

❷ **Inference: Handle Large Traffic Volume on Commodity Hardware.** Operators often manage high traffic volumes (e.g., 10Gbps to 100Gbps) in enterprises, requiring real-time systems that handle flows that often exceed model inference rates. Previous solutions address this by either heavily downsampling the traffic or relying on specialized hardware, such as customized SmartNICs, which constrains the range of models they can support. These methods either perform per-packet inference, resulting in repetitive computations for flow-level analysis, or buffer 4 to 100 packets per flow, which introduces prohibitive latency. In response, `ServeFlow` [7] is introduced as a model-serving solution tailored for network data analysis tasks. `ServeFlow` dynamically selects the number of packets to process and the appropriate model for each flow, optimizing latency, service rate, and accuracy. It assigns flows to slower models only when the fastest model's output is inadequate. `ServeFlow` can infer 76.3% of flows in under 16 ms, achieving a 40.5x speed-up in median end-to-end serving latency, while maintaining high service rates and accuracy. Even with thousands of features per flow, `ServeFlow` processes over 48.5k new flows per second on a 16-core commodity CPU, matching the flow rates of city-level network backbones. To achieve an optimal balance between accuracy and speed, I also explored applying Bayesian Optimization to select the most relevant features [8] and employ adaptive model swapping techniques to dynamically switch between online models [9].

❸ **Monitoring and Adaptation: Deal with Evolving Networks.** Networks frequently experience changes due to operational and environmental shifts, such as equipment updates or evolving user patterns in cellular networks [11], leading to inaccuracies in model inferences. These inaccuracies are often caused by concept drift, where the relationship between features and the target variable shifts over time. Mitigating concept drift is crucial for the successful deployment of ML models in dynamic environments, such as Cellular networks. Our measurements [10] reveal that concept drift impacts several key performance indicators (KPIs), regardless of the model type, training data size, or temporal window. Frequent retraining with newly available data is insufficient to mitigate concept drift and can even worsen model performance. Local Error Approximation of Features (`LEAF`), is a novel framework designed to combat concept drift [10]. `LEAF` works by identifying drift, pinpointing which features and time periods contribute most to it, and applying mitigation strategies such as feature forgetting and over-sampling. `LEAF` is evaluated against industry-standard approaches (e.g., periodic retraining) using over four years of cellular KPI data. Tests with Verizon, a major U.S. cellular provider demonstrate that `LEAF` consistently outperforms both periodic and event-triggered retraining on complex, real-world data, while also reducing the frequency and cost of retraining operations.

❹ **Applications: Downstream Tasks in Security and Privacy.** Through developing ML solutions for network operations across various devices, I have become interested in the impact of network systems on end-users, especially on security and privacy. I use a cross-layer approach to analyze data from physical and network layers to derive insights at the application layer, offering security and privacy solutions.

Securing Network Operations Against Spoofing Attacks. Network infrastructures, particularly those relying on GPS for synchronization and routing, are vulnerable to wireless spoofing attacks that can disrupt operational integrity. I investigated sophisticated attack techniques [13, 14] capable of stealthily manipulating GPS signals, potentially leading

to network misconfigurations and service disruptions. To counteract these threats, I developed a cost-effective detection strategy [12] that can be integrated into existing network systems. By analyzing signal angles of arrival (AoAs) and cross-referencing them with known satellite constellations, this method enables network operators to detect spoofing attacks with 95%-100% accuracy in just five seconds, thereby enhancing the security and reliability of network operations without significant performance overhead.

Privacy Preservation in Network Traffic Monitoring. Network operators monitor traffic for security and performance, raising privacy concerns for end-users. I introduced `Privacy Plumber` [15], a system that visualizes data flows and privacy risks using augmented reality and traffic analysis. `Privacy Plumber` overlays information about uplink traffic and its privacy impact within the physical view of devices, enabling informed decisions, such as scheduled blocking. A user study with six participants showed increased awareness and proactive privacy decisions.

Ongoing and Future Work

Fine-grained insight and control over networked interactions are essential to ensuring reliability and resilience in our increasingly interconnected world. My vision is to develop automated systems that *empower all stakeholders to interpret and manage networked interactions with reliability and transparency.* My current research has laid the groundwork by designing operational networked machine learning systems. I will expand the ambit of my research to address different stakeholders—model developers, network operators, and network service users—in both machine-to-machine and machine-to-human interactions, especially in critical areas like networking and cybersecurity. To achieve my vision, I will pursue three key lines of research:

Operationalize Synthetic Network Data. My vision in this direction is to innovate and systematize the opportunities of synthetic network data to enhance ML-driven network systems. This approach will enable network analytical providers to integrate synthetic data techniques—such as emerging deep generative models, data partitioning, and targeted data generation—through clear, deployable workflows. As traffic volumes, storage costs, data boundaries, and retention/compliance requirements increase, network operators face growing challenges in retaining network data. My ongoing and future efforts focus on developing deployable methods for generating and leveraging synthetic network data systematically. These methods aim to alleviate recollection, labeling, and storage burdens while maintaining the utility of the data for downstream ML tasks. (1) One key direction is data augmentation in unseen conditions through synthetic traces. By generating realistic synthetic data that "mixes" existing conditions, datasets can be expanded in what-if scenarios. This approach also enhances the accuracy and robustness of network models by introducing controlled variability, reducing dependence on collecting large, raw datasets. (2) Additionally, orchestrating synthetic data for cross-silo AI pools insights from multiple decentralized sources. It enables collaborative model training across different networks. By treating generative models as proxies for data across silos, it improves model generalization while preserving data privacy. However, we must carefully choose which silos to include: when does adding more synthetic data improve model outcomes, and when does it hinder them in real-world applications? (3) Another area of focus is longitudinal network data analysis. Enterprise data governance usually requires the implementation of a retention policy, but gaining network data over extended periods is crucial for tasks such as trend analysis, performance monitoring, and anomaly detection. Deep generative models that accurately reflect long-term network behavior can reduce storage needs while still providing valuable insights into network dynamics over time.

Evolve from Monolithic Models to Compound Model Networks. Building on the foundation of ServeFlow [7], my vision is to transition from singular fast-slow architectures to expansive networks composed of tens or hundreds of hierarchically differentiated models. These models will operate across a wide range of time scales—some analyzing data in sub-millisecond intervals, others adapting over extended periods—to provide insights at varying levels of granularity. By integrating models with diverse temporal granularities, and architectures, I aim to enable prompt and informed actions while effectively balancing speed, accuracy, and resource utilization.

This evolution involves (1) developing sophisticated hand-off strategies to determine when and where to delegate tasks between models, (2) exploring dynamic routing algorithms to select optimal model pathways based on real-time conditions, and (3) identifying optimal model placement within ISPs or data centers to enhance performance and reduce latency. (4) Additionally, implementing expressive mesh network policies will enhance modularity, adaptability, and fault tolerance, allowing the network to reconfigure in response to changing workloads or node failures. By abstracting models as interconnected network components, I aim to establish principled design methodologies for compound model networks. This approach will unlock new efficiencies in resource management, scalability, and system robustness across various applications, such as traffic analysis, video analytics, and large language model serving.

Leverage Cross-Layer Traffic Analysis for Emerging Applications. As ML techniques for traffic analysis become increasingly accurate and powerful, they offer new opportunities that extend beyond traditional network optimization and security. By harnessing cross-layer insights—where network and transport layer data inform interpretations at the application layer—we can analyze network behaviors that reflect or impact human activities in various contexts. Other signals often have network traffic companions, and even when events occur high up the stack, analyzing underlying network data can yield valuable inferences about application-level activities or human behaviors. This cross-stack analysis enables us to interpret subtle, context-driven patterns from network data, providing actionable triggers for real-world problems. For instance, in detecting intimate partner violence, perpetrators may exploit their intimate access to bypass security measures on victims' smartphones. By combining network traffic patterns with other modalities, we can identify unusual behaviors that signal privacy breaches, offering an automated and scalable solution to a growing issue. Another potential application is managing whole-house screen time across heterogeneous hardware by analyzing data from various devices in a privacy-preserving manner. These examples highlight how fine-grained network inference leverages traffic patterns as a valuable source of insight into complex, real-world challenges affecting human life.

References

[1] Raouf Boutaba, Mohammad A Salahuddin, Noura Limam, Sara Ayoubi, Nashid Shahriar, Felipe Estrada-Solano, and Oscar M Caicedo. A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, 9(1):1–99, 2018.

[2] Robin Sommer and Vern Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *Proceedings of the IEEE symposium on security and privacy (Oakland)*, pages 305–316. IEEE, 2010.

[3] Xi Jiang*, **Shinan Liu***, Aaron Gember-Jacobson, Paul Schmitt, Francesco Bronzino, and Nick Feamster. Generative, high-fidelity network traces. In *ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets)*, Cambridge, Massachusetts, 2023.

[4] Xi Jiang, **Shinan Liu**, Aaron Gember-Jacobson, Arjun Nitin Bhagoji, Paul Schmitt, Francesco Bronzino, and Nick Feamster. Netdiffusion: Network data augmentation through protocol-constrained traffic generation. In *Proceedings of the ACM on Measurement and Analysis of Computer Systems (SIGMETRICS)*, pages 1–14, Venice, Italy, 2024.

[5] Andrew Chu, Xi Jiang, **Shinan Liu**, Arjun Bhagoji, Francesco Bronzino, Paul Schmitt, and Nick Feamster. Feasibility of state space models for network traffic generation. In *Proceedings of the 2024 SIGCOMM Workshop on Networks for AI Computing (NAIC)*, pages 9–17, 2024.

[6] **Shinan Liu**, Tarun Mangla, Ted Shaowang, Jinjin Zhao, John Paparrizos, Sanjay Krishnan, and Nick Feamster. Amir: Active multimodal interaction recognition from video and network traffic in connected environments. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT/UbiComp)*, Cancun, Mexico, 2023.

[7] **Shinan Liu**, Ted Shaowang, Gerry Wan, Jeewon Chae, Jonatas Marques, Sanjay Krishnan, and Nick Feamster. Serveflow: A fast-slow model architecture for network traffic analysis. In *Submission*, 2024.

[8] Gerry Wan, **Shinan Liu**, Francesco Bronzino, Nick Feamster, and Zakir Durumeric. Cato: End-to-end optimization of ml-based traffic analysis pipelines. In *Submission*, 2024.

[9] Xi Jiang, **Shinan Liu**, Saloua Naama, Francesco Bronzino, Paul Schmitt, and Nick Feamster. Ac-dc: Adaptive ensemble classification for network traffic identification. In *Submission.*, 2023.

[10] **Shinan Liu**, , Francesco Bronzino, Paul Schmitt, Arjun Nitin Bhagoji, Nick Feamster, Hector Garcia Crespo, Timothy Coyle, and Brian Ward. Leaf: Navigating concept drift in cellular networks. In *Proceedings of the ACM SIGCOMM International Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, pages 1–12, Paris, France, 2023.

[11] **Shinan Liu**, Paul Schmitt, Francesco Bronzino, and Nick Feamster. Characterizing service provider response to the covid-19 pandemic in the united states. In *Proceedings of the Passive and Active Measurement Conference (PAM)*, pages 20–38, Brandenburg, Germany, 2021.

[12] **Shinan Liu***, Xiang Cheng*, Hanchao Yang, Yuanchao Shu, Xiaoran Weng, Ping Guo, Kexiong Curtis Zeng, Gang Wang, and Yaling Yang. Stars can tell: a robust method to defend against gps spoofing attacks using off-the-shelf chipset. In *Proceedings of the USENIX Security Symposium (USENIX Security)*, pages 3935–3952, 2021.

[13] Kexiong Curtis Zeng, Yuanchao Shu, **Shinan Liu**, Yanzhi Dou, and Yaling Yang. A practical gps location spoofing attack in road navigation scenario. In *Proceedings of the International Workshop on Mobile Computing Systems and Applications (HotMobile)*, pages 85–90, 2017.

[14] Kexiong Curtis Zeng, **Shinan Liu**, Yuanchao Shu, Dong Wang, Haoyu Li, Yanzhi Dou, Gang Wang, and Yaling Yang. All your gps are belong to us: Towards stealthy manipulation of road navigation systems. In *Proceedings of the USENIX Security Symposium (USENIX Security)*, pages 1527–1544, 2018.

[15] Stefany Cruz, Logan Danek, **Shinan Liu**, Christopher Kraemer, Zixin Wang, Nick Feamster, Danny Yuxing Huang, Yaxing Yao, and Josiah Hester. Toward identifying home privacy leaks using augmented reality. In *Proceedings of the Symposium on Usable Security and Privacy (NDSS USEC)*, San Diego, CA, 2023.